Q&A

1. What was the project objective, and what steps did you take to achieve it?
For this project, we had two objectives: Identify key influencers affecting energy consumption and develop a model that accurately predicts energy consumption using historical data. To meet our objective, we processed the data, analyzed it, engineered features, and built models using linear regression and random forest regressors.

2. From your analysis, what were the primary factors impacting energy consumption?
Based on our EDA of the dataset, we found that the temperature of the building, HVAC usage, and occupancy were the top factors affecting energy consumption.

3. Can your model predict future energy consumption using historical data?
Yes, our model can predict energy usage by leveraging time-stamped energy consumption data and variables influencing energy consumption. However, our linear regression mode's R2 value of 0.64 indicates a low to moderate level of accuracy. Further improvement of the model's accuracy is needed before real-world implementation.

4. What factors do you believe contributed to your model's accuracy?
The dataset we pulled from Kaggle was artificially generated, which may limit the accuracy of the relationships and correlations in the data we used for model training. This could impact our model's learning capability and predictive output. In addition, our choice of model and feature selection could have influenced the output. More time and testing are needed for further refinement.

5. Why do you think the linear regression model outperformed the random forest regressor?
We've run multiple model training iterations for linear regression and random forest regressor. We concluded that linear regression was a better fit for our data, contrary to our initial hypothesis favoring the random forest regressor. This suggests that the relationships between our input features and the target variable may be more linear than anticipated.

6. Please give me a quick summary of your model training iterations.
We iteratively trained our models by adjusting the features, training/testing splits, and hyperparameters. We conducted seven iterations for the random forest regressor and five iterations for linear regression. Our objective was to find the optimal combination that provided the best performance based on the R2 metric.

7. What feature engineering tasks did you perform?
For feature engineering, we first extracted the AM/PM information from the timestamp column to improve the model's learning process, as the timestamp column only contained unique values. Then, we scaled numerical features and encoded categorical values to make the data more suitable for the model training process.

8. Tell me about your data visualizations.
We used histograms to understand feature distribution and box plots to search for outliers. Then, we used the correlation matrix and the feature importance bar graph to analyze the top influencers that affect energy consumption. Finally, we built plots comparing actual vs. predicted outputs and residuals, which helped our team identify model biases and accuracy/performance.

9. What recommendations will you be making based on your project's outcome?
Based on the project outcome, we recommend installing smart thermostats or HVAC units capable of adjusting settings based on top factors like the weather forecast (temperature), occupancy, and HVAC usage to optimize energy usage. To build this device, we recommend using our energy consumption model as the predictive engine that informs the device when and how to adjust the settings effectively.

10. What are the next steps for the model?
Our next step is very straightforward. We plan to enhance the model's accuracy by exploring more complex models, performing further feature engineering, and using real-world data to capture the relationships between the features better.